

Cited Reference ↗

METHOD AND DEVICE FOR FILTERING SPAM MAIL BY RECOGNIZING MASS OF MAILS

Publication number: KR20030069567

Publication date: 2003-08-27

Inventor: KIM GYEONG TAE (KR); KIM YEONG JUN (KR); NAM SE DONG (KR); SHIN JUNG HO (KR)

Applicant: NEOWIZ CO LTD (KR)

Classification:

- **international:** (IPC1-7): G06F17/60

- **European:**

Application number: KR20020009412 20020222

Priority number(s): KR20020009412 20020222

[Report a data error here](#)

Abstract of KR20030069567

PURPOSE: A method and device for filtering spam mails by recognizing mass of mails is provided to block and manage the spam mails effectively. **CONSTITUTION:** A mail managing unit(210) supplies a mail being received from an exterior in a mail storage(220), transmits a received mail to a designated user, and selects a spam mail out of received mails in accordance with a spam mail analysis result of a spam mail analysis unit(250). The mail storage(220) records a tag with respect to the spam mail selected in the mail managing unit(210), a corresponding spam mail is displayed, and checks a stored spam mail in accordance with a request of the user. A mass mail analysis unit(230) extracts the 'n' number of characters which are used frequently out of characters included in each mail received in the mail managing unit(210) as a representative character string, classifies similar mails having the same representative character string as the extracted representative character string. If the number of persons who receive similar mails having the identical representative character string is more than a predetermined number, the mass mail analysis unit(230) judges the mails as a mass of mails. A non-similar mail analysis unit(240) excludes a mail having an identical representative character string and different contents from a similar character string. The spam mail analysis unit(250) compares a keyword of a corresponding mail with a predetermined keyword, analyzes a spam mail, and supplies a spam mail analysis result to the mail managing unit(210).

Data supplied from the **esp@cenet** database - Worldwide

Cited Reference 乙

특2003-0069567

(19) 대한민국특허청(KR)

(12) 공개특허공보(A)

(51) Int. Cl.⁷
G06F 17/60D0(11) 공개번호 특2003-0069567
(43) 공개일자 2003년08월27일

(21) 출원 번호 10-2002-0009412

(22) 출원 일자 2002년02월22일

(71) 출원인 주식회사 네오위즈

서울특별시 강남구 삼성1동 아셈타워

(72) 발명자 신중호

서울특별시 강남구 일원동 우성7차아파트 114-705호

남세동

서울특별시 강남구 도곡동 467-18 우성캐릭터빌 816호

김영준

충청남도 천안시 성황동 5-27

김경태

서울특별시 서초구 잠원동 66-3 동아아파트 106동 205호

(74) 대리인 이철희

심사청구 : 있음

(54) 대량 메일의 파악에 기반한 스팸 메일 필터링 방법 및 장치

요약

본 발명은 인터넷을 통하여 무차별적으로 살포되는 대량의 광고성 및 상업성 메일을 수신한 사용자의 충수에 기반하여 대량 메일을 분류하고, 분류된 대량 메일에 기반하여 스팸 메일을 인식하는 스팸 메일 필터링 방법 및 장치에 관한 것이다. 본 발명은 외부로부터 수신된 각각의 메일에 포함된 문자 중에서 사용빈도수가 높은 N (N 은 자연수) 개의 문자를 상기 메일을 대표하는 대표 문자열로 추출하고, 상기 대표 문자열이 동일한 메일들의 개수를 상기 대표 문자열로 누적하고, 상기 대표 문자열로 누적된 메일들의 누적 개수가 기설정 개수 이상인 메일들을 대량으로 발송된 유사 메일로 분석하고, 상기 유사 메일로 분류된 각각의 메일에 포함된 키워드와 기설정 키워드를 비교하여 기설정 사전 설정치 이상의 유사도를 갖는 메일을 상기 스팸 메일로 인식한다. 또한, 본 발명은 유사 메일로 분류된 메일들 중에서 대표 문자가 동일하지만 메일의 내용이 일부만 바뀐 메일도 유사 메일로 판단하여 스팸 메일로서 인식할 수 있고, 또한, 대표 문자가 동일한 메일들 중에서 우연히 대표 문자가 같지만 그 내용이 유사하지 않은 메일을 제외시킴으로써 스팸 메일로서의 인식률의 정확도를 높일 수 있다.

대표도

도2

색인어

스팸 메일, 필터링, 내용 기반 유사 메일, 대표 문자, 형태소 분석

명세서

도면의 간단한 설명

도 1은 메일 서비스 시스템의 블럭 구성도.

도 2는 본 발명에 따른 메일 서버의 블럭 구성도,

도 3은 대량 메일 분석부의 상세 블럭 구성도,

도 4는 비유사 메일 분석부의 상세 블럭 구성도,

도 5는 스팸 메일 분석부의 상세 블럭 구성도,

도 6 및 도 7은 각각 스팸 메일로 인식된 메일의 사후 처리 과정을 설명하는 흐름도이다.

<도면의 주요 부분에 대한 부호의 설명>

| | |
|------------------|------------------|
| 110 : 외부 메일 서버 | 130 : 메일 서버 |
| 210 : 메일 관리부 | 230 : 대량 메일 분석부 |
| 240 : 비유사 메일 분석부 | 250 : 스팸 메일 분석부 |
| 310 : 대표 문자 추출부 | 320 : 동일 메일 분류부 |
| 330 : 대량 메일 판별부 | 410 : 문자 빈도수 계산부 |
| 420 : 빈도수 평균 계산부 | 430 : 빈도수 평균 비교부 |
| 510 : 형태소 분석부 | 520 : 키워드 추출부 |
| 540 : 스팸 메일 인식부 | 550 : 스팸 용어 저장부 |

발명의 상세한 설명

발명의 목적

발명이 속하는 기술분야 및 그 분야의 종래기술

본 발명은 인터넷을 통하여 무차별적으로 배포되는 스팸 메일(Spam Mail)을 필터링하는 방법 및 장치에 관한 것으로, 보다 상세하게는, 동일한 메일을 수신한 사용자의 수에 기반하여 대량 메일을 분류하고, 분류된 대량 메일에 기반하여 스팸 메일을 필터링하는 방법 및 장치에 관한 것이다.

인터넷 메일 서비스에 있어서, 가장 큰 문제점 중의 하나는 인터넷을 통하여 무차별적으로 배포되는 광고성 및 상업성 전자 메일(E-mail), 이하 간략히 메일이라고 함)을 들 수 있다. 보통 스팸 메일이라고 불리우는 광고성 및 상업성 메일은 광의적 의미로서 사용자가 읽을 필요가 없는 메일, 즉 사용자가 해당 메일의 내용을 보지 않고 삭제하기를 원하는 메일을 의미하며, 협의적 의미로는 사용자의 수신 의사와 무관하게 통신이나 인터넷을 통해 무차별적으로 대량 살포되는 메일을 의미한다.

많은 사용자들은 스팸 메일이라고 불리우는 대량 살포되는 광고성 및 상업성 메일로 인하여 본인 의사와 무관하게 불필요한 메일을 수신하고, 불필요한 메일을 지워야하는 관리적인 불편을 겪고 있으며, 불필요한 스팸 메일로 인하여 전체 인터넷의 트래픽이 증가하는 등 막대한 비효율성을 수반하고 있다.

인터넷 메일 서비스를 통하여 수신되는 스팸 메일을 필터링하는 방법은 다음과 같이 여러 가지 방법이 있다.

첫 번째 방법은 스팸 메일을 보내는 발신자측의 메일 주소(E-mail Address)를 리스트로 관리하여, 그 리스트에 속한 메일 주소로부터 오는 메일을 스팸 메일로 간주하는 방법이다. 다시 말해서, 이 방법은 메일 서버에서 스팸 메일을 주로 보내는 메일 주소를 조사하여, 수신 거부 리스트를 만들고, 해당 주소에서 보내진 메일들을 스팸 메일로 인식하는 것이다. 그러나, 이러한 리스트 관리 방법은 스팸 메일 여부를 판단하고, 해당하는 메일 주소를 일일이 수동으로 관리하는 수작업이 요구된다는 점과, 일단 스팸 메일을 발송한 메일 주소가 리스트에 등록되면 해당하는 주소에서 발송되는 모든 메일이 스팸 메일로 취급되어 간혹 있을 수 있는 중요한 메일조차도 열람하지 못하는 경우가 발생된다는 단점이 있다. 또한, 스팸 메일을 작성하여 발송할 때 발신 주소를 기준의 주소로 하지 않고 발신자 주소를 임의로 생성하거나 보낼때마다 변경하여 스팸 메일을 발송하는 경우에는 스팸 메일의 차단 기능이 거의 무력화되는 단점이 있다.

두 번째 방법은 수신된 메일의 제목이나 내용을 분석하여, '광고', '돈벌기', '홍보' 등과 같이 광고나 상업적인 문구의 내용이나 홍보성 문구의 내용을 나타내는 특정 단어를 사전 설정치 이상 포함하는 경우 스팸 메일로 인식하는 방법이다. 이 방법은 첫 번째 방법과 달리 스팸 메일을 보낸 사람의 주소를 기준으로 하지 않고, 메일 내용을 분석하여 스팸 메일 여부를 판단한다는 차별성이 있다. 그러나, '광고'나 '돈벌기'와 같은 특정 단어를 포함하는 메일 중에서 스팸 메일이 아닌 많은 경우를 스팸 메일로 인식하는 오류가 있고, 스팸 메일의 특성을 나타내는 모든 단어를 분석하는 것이 현실적으로 불가능하다는 단점이 있다.

세 번째 방법은 사회적 필터링(Social Filtering) 기법을 이용한 것으로, 사용자들이 수신한 메일에 대하여, 스팸 메일 여부를 메일 서버 혹은 스팸 메일 관리 서버에 신고하여 스팸 메일을 필터링하는 것이다. 메일 서버는 사전 설정치 이상의 사용자들로부터 스팸 메일이라고 신고된 메일에 대하여, 같은 메일을 수신한 다른 사용자들에게도 해당 메일이 스팸 메일임을 알린다. 이 방법은 스팸 메일 여부를 판단하는 기준을 사용자들의 해당 메일에 대한 스팸 메일 판단의 통계에 기반하는 방법이므로, 전술한 두 번째 방법과 비교하여 사용자 관점에서 비교적 정확한 스팸 메일 판단이 가능하다는 장점이 있다. 그러나, 이 방법의 단점은 사용자들의 스팸 메일에 대한 판단 정보가 충분하게 확보되지 못한 메일에 대해서는 정확하게 스팸 메일인지의 여부를 판단할 수 없다는 점과, 각 메일에 대하여 사전 설정치 이상의 충분한 사용자들의 스팸 메일의 신고 데이터가 축적되기 전까지는 해당 메일이 스팸 메일인지의 여부를 신속하게 판단할 수 없다는 단점이 있다.

한편, 스팸 메일의 필터링을 위한 스팸 메일 인식 방법에는 동일 메일 인식 방법이 있다. 동일 메일 인식 방법은 특정 메일을 스팸 메일이라고 판단한 후, 그 스팸 메일과 동일한 메일을 수신한 사용자들에게 해당하는 메일을 스팸 메일로 처리하기 위하여 적용된다. 통상적인 동일 메일 인식 방법은 메일을 보낸 사람의 주소, 메일의 제목 및 메일의 내용과 같은 정보 혹은 이를 정보들의 조합을 조사하여, 같은 내용을 가지면 동일한 메일로 인식한다.

그러나, 동일 메일 인식 방법은 전술한 기준의 방법들과 같이 보낸 사람의 주소와 제목 등으로 동일 메일 여부를 판단할 경우, 메일의 전체 내용은 동일하지만 메일의 내용 중에서 특정 부분이 변하는 유사 메일 예를 들면, '홍길동 귀하'와 같이 수신자만이 달라지는 메일을 동일 메일로 인식하지 못한다는 한계가 있

다. 더욱이, 내용이 같은 메일이라도 보낸 사람의 주소가 변경된 경우에도 동일 메일로 인식하지 못하는 문제점이 있다.

발명이 이루고자 하는 기술적 과제

그러므로, 본 발명은 스팸 메일을 효율적으로 차단하고 관리할 수 있는 스팸메일 필터링 방법 및 장치를 제공하는 것을 그 목적으로 한다.

본 발명의 다른 목적은 내용이 유사한 다수의 메일을 대량 메일로 판단하고, 대량 메일의 판단에 기반하여 스팸 메일을 필터링하는 방법 및 장치를 제공하는 것이다.

발명의 구성 및 작용

전술한 목적을 달성하기 위한 본 발명의 바람직한 실시예에 따르면, 메일 서버에서 스팸 메일을 필터링하는 방법은, 외부로부터 수신된 각각의 메일에 포함된 문자 중에서 사용 빈도수가 높은 N(N은 자연수) 개의 문자를 상기 메일을 대표하는 대표 문자열로 추출하는 대표 문자 추출 단계; (b) 상기 대표 문자열이 동일한 메일들의 개수를 상기 대표 문자열별로 누적하는 메일 누적 단계; (c) 상기 대표 문자열별로 누적된 메일들의 누적 개수가 기설정 개수 이상인 메일들을 대량 메일로 분석하는 대량 메일 분석 단계; 및 (d) 상기 대량 메일로 분류된 각각의 메일에 포함된 복수개의 키워드와 기설정 키워드를 비교하여 기설정 사전 설정치 이상의 유사도를 갖는 메일을 상기 스팸 메일로 인식하는 스팸 메일 분석 단계를 포함하는 것을 특징으로 한다.

본 발명의 또 다른 실시예에 따르면, 상업성 또는 광고성 스팸 메일을 필터링하는 장치는, 외부로부터 수신된 메일들 중에서 스팸 메일 인식 결과에 따라 상기 수신된 각각의 메일에 대한 스팸 메일의 여부를 사용자에게 알려주는 메일 관리부; 및 상기 메일 관리부에서 수신된 각각의 메일에 포함된 내용을 분석하여 내용이 유사한 메일들을 분류하고, 분류된 유사 메일들이 기설정 개수 이상으로 대량인 유사 메일들을 상기 스팸 메일로 분석하는 대량 메일 분석부를 포함하는 것을 특징으로 한다.

본 발명의 또 다른 실시예에 따르면, 상업성 또는 광고성 스팸 메일을 필터링하는 장치는, 외부로부터 수신되는 각각의 메일을 스팸 메일 인식 결과에 따라 스팸 메일로서 관리하는 메일 관리부; 상기 메일 관리부에서 수신된 각각의 메일에 포함된 문자 중에서 사용 빈도수가 높은 N(N은 자연수) 개의 문자를 상기 메일을 대표하는 대표 문자로 추출하고, 상기 추출된 대표 문자가 동일한 메일들이 사전 설정치 이상인 메일들을 대량 메일로 분류하는 대량 메일 분석부; 및 상기 대량 메일로 분류된 각각의 메일에 포함된 단어들 중에서 그 메일을 대표하는 키워드와 기설정 스팸성 키워드를 비교하여 사전 설정치 이상의 유사도를 갖는 메일을 스팸 메일로 인식하고 상기 메일 관리부로 상기 스팸 메일 인식 결과를 제공하는 스팸 메일 분석부를 포함하는 것을 특징으로 한다.

이하, 본 발명의 바람직한 실시예를 첨부된 도면을 참조하여 다음과 같이 상세히 설명한다.

본 발명의 상세한 설명에 앞서, 본 발명과 연관된 용어에 대하여 정의하면 다음과 같다.

* 대량 메일 : 사전 설정치 이상의 수신자에게 발송된 메일의 내용이 유사한 메일

* 스팸 필터링 : 메일 사용자에게 유용하지 않은 무차별적 광고성 메일을 구분하여, 스팸 메일의 여부를 표시하고, 스팸 메일을 자동으로 별도의 보관함으로 이동시키는 등 향후 유사 메일의 수신을 차단하는 방법 또는 서비스

* 동일 메일 : 내용이 100 % 동일한 메일

* 유사 메일 : 메일의 내용이 사전 설정치 이상 동일한 메일

* 스팸메일 :

1) 광의적 정의 : 사용자가 읽을 필요가 없는 메일. 즉 사용자가 해당 메일의 내용을 보지도 않고 삭제하여도 무방하다고 여겨지는 메일

2) 협의적 정의 : 사용자의 수신 의사와 무관하게 통신이나 인터넷을 통해 무차별적으로 대량 살포되는 광고성 및 상업성 메일

* 스팸성 단어 : '광고', '구매' 또는 '찬스' 등 광고성 메일임을 인식할 수 있는 단어

이제, 도 1을 참조하면, 본 발명에 따라서 구성된 스팸 메일 필터링 시스템의 개략적인 블럭 구성이 도시된다.

외부 메일 서버(110)는 외부의 발신자에 의해 작성된 메일을 인터넷(120)을 통하여 본 발명의 메일 서버(130)로 송신하며, 메일 서버(130)는 외부 메일 서버(110)로부터 송신된 메일을 수신하고, 수신된 메일을 그 메일에서 지정하는 사용자(140)들에게 전달한다.

또한, 메일 서버(130)는 수신된 메일 중에서 내용이 유사한 메일을 접수한 사용자(140)의 수를 기준으로 메일이 대량으로 발송되었는지를 판단하여 스팸 메일을 판단하는 기준으로 삼는다. 대량 메일을 판단하는 과정에서는 각각의 메일의 내용이 사전 설정치 이상의 유사도를 가지는 경우 유사한 메일로 간주하는 내용 기반의 유사 메일 판단 방법을 적용한다. 메일 서버(130)는 유사 메일 판단 기준에 속한 메일 중에서 광고나 상업적인 스팸성 단어를 가지고 있는 메일을 스팸 메일로 규정하고, 스팸 메일로 규정한 메일을 별도로 관리 또는 폐기하거나 해당 메일을 수신한 각 사용자들에게 전송된 메일이 스팸 메일임을 알려준다.

사용자(140)는 웹 브라우저가 내장된 컴퓨터 또는 그 컴퓨터의 사용자를 의미한다. 사용자(140)는 메일 서

버(130)로부터 스팸 메일의 통보에 따라 수신된 메일을 폐기할 수 있다.

도 2는 도 1에 도시된 본 발명에 따른 스팸 필터링 메일 서버(130)의 상세 블럭 구성도를 도시한다.

메일 서버(130)는 메일 관리부(210), 메일 저장부(220), 대량 메일 분석부(230), 비유사 메일 분석부(240) 및 스팸 메일 분석부(250)를 구비한다.

메일 관리부(210)는 외부로부터 수신되는 메일을 메일 저장부(220)에 제공하여 저장하도록 하고, 수신된 메일을 지정된 사용자(140)에게 전달하며, 스팸 메일 분석부(250)의 스팸 메일 분석 결과에 따라 수신된 메일들 중에서 스팸 메일을 선별하는 기능을 수행한다. 메일 관리부(210)는 스팸 메일 분석부(250)에 의해 분석된 결과에 따라 스팸 메일로 분석된 메일을 제외한 메일들만을 사용자(140)에게 전송하거나, 수신된 모든 메일을 그대로 사용자(140)에게 제공한 후 스팸 메일 분석부(250)에 의해 분석된 결과를 사용자에게 통보할 수도 있다.

메일 저장부(220)는 메일 관리부(210)에서 선별된 스팸 메일에 대하여 태크(Tag)를 기록함으로써 해당하는 메일이 스팸 메일임을 나타내며, 사용자(140)의 요구에 따라 저장된 스팸 메일을 확인시켜준다.

대량 메일 분석부(230)는 메일 관리부(210)에서 수신한 각각의 메일에 포함된 문자 중에서 사용 빈도수가 높은 N(N은 자연수) 개의 문자를 그 메일을 대표하는 대표 문자열로 추출하고, 추출된 대표 문자열과 동일한 대표 문자열을 갖는 유사 메일들을 분류하고, 동일한 대표 문자열을 갖는 유사 메일을 수신한 사용자의 수가 일정 사용자 이상일 때 그 메일들을 대량 메일이라고 판단한다. 대량 메일 분석부(230)에서 대량 메일을 판단하는 과정은 도 3을 참조하여 상세히 설명한다.

비유사 메일 분석부(240)는 대량 메일 분석부(230)에 의해 분석된 대량의 유사 메일들 중에서 대표 문자열이 동일하지만 메일 본문의 내용이 유사하지 않은 메일을 유사 메일들에서 제외시키는 기능과 더불어 대표 문자열이 동일하되 메일 본문의 내용이 일부만 변경된 메일을 유사 메일로서 인식하는 기능을 수행한다. 비유사 메일 분석부(240)의 동작은 도 4를 참조하여 상세히 설명한다.

스팸 메일 분석부(250)는 대량 메일 분석부(230)에 의해 분석된 대량 메일을 기준으로 해당하는 메일에 포함된 키워드를 기설정 키워드와 비교하여 싱호 일치하는 메일을 스팸 메일로 분석하고 메일 관리부(210)로 스팸 메일 분석 결과를 제공한다. 스팸 메일 분석부(230)의 동작은 도 5에서 상세히 설명한다.

도 3은 도 2에 도시된 대량 메일 분석부(230)의 상세 블럭 구성을 도시한다. 대량 메일 분석부(230)는 대표 문자 추출부(310), 동일 메일 분류부(320), 메일 관리 테이블(330) 및 대량 메일 판별부(340)를 구비한다.

대표 문자 추출부(310)는 수신된 메일에 포함된 문자 및 각 문자의 빈도수. 즉 사용 회수를 계산한다. 예컨대, 수신된 메일이 'EAAEAB AD CD EEAD'이라는 내용을 담고 있는 경우, 메일에서 사용된 문자들은 모두 'A B C D E'이다. 이때, 이를 각 문자의 사용 빈도수는 각각 '5 1 1 3 4'이다. 여기서, 메일에서 사용된 문자들 중에서 사용 빈도수가 가장 많은 순서대로 N 개의 문자를 대표 문자열로 추출한 다음, 추출된 N 개 문자 각각의 사용 빈도수를 각 문자의 코드로 부여하여 N 개의 코드로 구성된 대표 코드열을 생성한다. 예를 들면, N = 3인 경우에는 'A E D'가 그 메일을 대표하는 대표 문자열로서 선택되고, 메일의 내용을 대표하는 대표 문자열, 'A E D'에서 각 문자의 사용 빈도수, 즉, '5 4 3'를 해당 메일의 대표 코드열로서 추출한다. 이 때, 각 문자에 대응적으로 부여된 코드는 각 문자의 사용 빈도수로부터 첫째 자릿수만을 추출하여 사용한다. 물론, 메일에서 출현하는 각 문자의 사용 빈도수가 10 단위 또는 100 단위로 될 수도 있고, 그 단위를 그대로 대표 코드열에 사용할 수도 있지만, 대표 코드열의 간략화를 기하기 위하여 첫째 자릿수만을 추출하여 사용하는 것이 바람직하다.

본 발명에 있어서, 대표 코드를 생성하기 위하여 대표 문자를 이용하는 방법을 기술하고 있지만, 이와 달리 해쉬 함수(Hash Function)를 이용한 내용 요약에 기반하여 대표 코드를 생성하는 방법을 적용할 수도 있다. 내용 요약을 통하여 대표 코드를 생성하는 해쉬 함수 기반의 방법 중에서 대표적인 방법은 MD5(Message Digest5) 알고리즘을 들 수 있다. MD5 알고리즘은 어떠한 메시지의 내용을 요약(Digestion)하여 128 비트 해쉬 코드로 암호화하는 방법이다. 본 발명에서 MD5 알고리즘을 적용하기 위해서는 수신된 각각의 메일의 내용을 요약하여 128 비트 해쉬 코드를 생성하고, 생성된 128 비트 해쉬 코드를 본 발명에 서와 같이 동일 메일을 분류하기 위한 대표 코드값으로 사용할 수도 있다.

동일 메일 분류부(320)는 수신되는 각각의 메일에서 동일한 대표 코드열을 갖는 메일들을 분류한다.

메일 관리 테이블(330)은 RAM과 같은 메모리로 구현될 수 있으며, 동일 메일 분류부(320)에 의해 분류된 메일들을 대표 코드열별로 누적한다. 하기 표 1은 동일 메일 분류부(330)에 의해 분류된 메일들의 개수를 대표 코드열별로 누적하는 예를 나타낸다.

[표 1]

| 대표 코드열 | 메일 개수 | 메일 ID |
|--------|-------|-------------------|
| 543 | 120 | 111, 134, 343.... |
| 123 | 20 | 3434, 434, 34.... |
| . | . | . |

표 1에서, 수신된 각 메일이 '543' 및 '123'이라는 대표 코드열을 가지고 있고, 대표 코드열이 동일한 메일이 각기 '120' 및 '20' 개씩 누적되어 있으며, 그 대표 코드열을 갖는 메일 ID가 리스트되어 있음을 알 수 있다.

대량 메일 판별부(340)는 메일 관리 테이블(330)을 참조하여 대표 코드열에 속하는 메일의 누적 개수가 기 설정 개수, 예컨대, 100 개 이상인 메일들을 대량 메일로 판별한다. 이것은 100 명 이상의 사용자(140)가 모두 내용이 유사한 메일을 수신한 것 또는 유사한 메일이 100 개 이상 발송된 것을 의미한다. 대량 메일 판별부(340)의 판별에 의하면, 표 1에서는 대표 코드열 '543'을 갖는 메일의 개수가 120으로서 기설정치를 초과하므로, 메일 ID가 111, 134 및 343인 메일들은 대량 메일로 판별된다.

본 발명에서는 내용이 유사한 메일을 수신한 사용자(140)의 수 또는 메일 발송 건수를 기준으로 대량 메일을 분류하고, 대량 메일로 분류된 메일은 다음에 설명하는 스팸 메일 분석부(250)에서 스팸 메일을 판단하는 기준이 된다.

물론, 대량 메일 판별부(340)에서 분류된 대량 메일을 그대로 스팸 메일로 간주하여 해당 메일을 수신한 사용자(140)에게 해당 메일이 스팸 메일임을 알려 줄 수도 있을 것이다. 또한, 대량 메일 판별부(340)에서 분류된 대량 메일을 그대로 스팸 메일로 간주하는 경우, 종래 기술의 첫 번째 방법에서 설명한 바와 같이, 스팸 메일을 보낸 발신자의 메일 주소를 리스트로 관리하고, 해당하는 메일 주소 리스트에 속한 메일 주소로부터 수신되는 메일을 모두 스팸 메일로 간주하여 조치를 취할 수도 있다. 또한, 대량 메일 판별부(340)에서 분류된 대량 메일을 그대로 스팸 메일로 간주하는 경우, 종래 기술의 두 번째 방법에서 설명한 바와 같이, 스팸 메일의 제목이나 내용을 검사하여 광고성 또는 상업성 특정 단어를 갖는 스팸 정보를 스팸 정보 리스트로 관리하여 그 특정 단어가 포함된 수신 메일을 모두 스팸 메일로서 필터링할 수도 있을 것이다.

도 4는 도 2에 도시된 비유사 메일 분석부(240)의 상세 블럭 구성이 도시된다. 비유사 메일 분석부(240)는 문자 빈도수 계산부(410), 빈도수 평균 계산부(420) 및 빈도수 평균 비교부(430)를 구비한다.

문자 빈도수 계산부(410)는 도 2의 대량 메일 분석부(230)에 의해 대량 메일로서 분석된 각각의 유사 메일에 포함된 전체 문자에 대한 빈도수를 계산한다.

빈도수 평균 계산부(420)는 문자 빈도수 계산부(410)에서 계산된 각 문자의 사용 빈도수의 평균을 계산한다. 빈도수 평균 계산부(420)에서 평균값을 계산하는 과정은 하기 표 2를 참조하여 설명한다.

(표 2)

| 유사 메일군 (A B C D E) | 사용 빈도수 | | | | |
|-----------------------|--------|-----|-----|-----|-----|
| | A | B | C | D | E |
| 메일 1 | 500 | 10 | 10 | 300 | 400 |
| 메일 2 | 510 | 10 | 10 | 300 | 400 |
| 메일 3 | 390 | 300 | 300 | 500 | 400 |
| . | . | . | . | . | . |

표 2에서는 세 개의 메일 1, 메일 2 및 메일 3이 도 2의 대량 메일 분석부(230)에 의해 유사 메일로 분류된 메일들로서, 각각의 메일이 모두 'A B C D E'라는 문자열을 가지고 있다고 가정한다. N = 3 인 경우 메일 1, 메일 2 및 메일 3의 대표 문자열은 각기 'A E D', 'A E D' 및 'D E A'로 다르지만, 모두 '5 4 3'이라는 동일한 대표 코드를 가지고 있다. 이들 메일들에서, 문자열 'A, B, C, D, E'의 빈도수 평균은 각기 $1400/3 (=466.66)$, $320/3 (=106.66)$, $320/3 (106.66)$, $1100/3 (366.66)$, $1200/3 (=400)$ 으로 계산된다.

빈도수 평균 비교부(430)는 빈도수 평균 계산부(420)에서 계산된 각 문자의 평균값과 빈도수를 벡터값으로 간주하여 비교하고, 이를 기의 벡터 유사도가 사전 설정치, 예컨대, 90 % ~ 95 %를 초과하지 않는 메일을 유사 메일군에서 제외시킨다. 이 과정에서 메일 3이 제외될 것이다.

본 발명에 있어서, 문서의 내용을 대표하는 벡터를 표현하기 위하여 문자 단위의 벡터값을 이용하는 방법을 기술하고 있지만, 이와 달리 단어 단위의 문서 내용을 대표하는 벡터값을 계산하는 방법을 적용할 수도 있다.

따라서, 비유사 메일 제외부(240)에 의해 진행되는 과정을 거치면서 우연히 대표 코드열은 같지만 메일의 내용은 유사하지 않은 메일을 비유사 메일로 판단하여 유사 메일에서 제외시킬 수 있다. 또한, 대표 코드열은 같은데 메일의 내용이 일부만 바뀐 메일, 예컨대, 수신자의 이름만 바뀐 메일의 경우에도 유사 메일로 판단할 수 있으므로, 본 발명에서 목적으로 하는 유사 내용을 갖는 메일을 유사 메일로서 인식하는 것이 가능하다. 빈도수 평균 비교부(430)에서 분석된 결과는 스팸 메일 분석부(250)로 제공된다.

도 5는 도 2에 도시된 스팸 메일 분석부(250)의 상세 블럭 구성이 도시된다. 스팸 메일 분석부(250)는 형태소 분석부(510), 키워드 추출부(520), 불용어 키워드 저장부(530), 스팸 메일 인식부(540) 및 스팸성 단어 저장부(550)를 구비한다.

형태소 분석부(510)는 대량 메일 분석부(230) 또는 비유사 메일 분석부(240)에서 대량 메일 또는 유사 메일로 분석된 각각의 메일에 포함된 문자열에 대하여 형태소 단위, 예컨대, 단어 단위의 형태소 분석을 수행한다. 형태소 분석은, 예를 들어 설명하면, '공짜로 쉬운 돈벌기 방법' 같은 구문에 대하여 '공짜(명사)+로(조사)+쉽(형용사)+은(어미)+돈벌기(명사)+방법(명사)'와 같은 방식으로 형태소 단위의 분석을 수행하는 과정이다. 형태소 분석부(510)에 의해 분석된 명사나 명사구 등의 복수개의 단어를 추출하고, 추출된 복수개의 단어를 각기 키워드 가능 후보로서 키워드 추출부(520)로 제공한다.

키워드 추출부(520)는 형태소 분석부(510)에서 제공된 키워드 가능 후보들 중에서 매일의 내용을 대표하는 키워드를 선택한다. 이때, 키워드를 선택하는 기준은 키워드 가능 후보들 중에서 불용어 키워드 저장부(530)에 수록되지 않은 모든 키워드 가능 후보를 키워드로서 추출한다. 불용어 키워드 저장부(530)는 색인 어로서의 가치가 적은 대명사, 관형사, 부사, 감탄사, 그리고 자주 출현하는 용언의 어간이 수록되어 있고, 또한 명사 중에서도 일반적으로 색인어로서의 가치가 희박한 것도 불용어로 수록한 불용어 사전으로서 사용된다. 이러한 불용어 키워드 저장부(530)는 ROM과 같은 메모리 소지로서 구현될 수 있다. 키워드 추출부(520)에서 키워드를 선택하는 방법으로서 불용어 사전을 이용하는 것으로 설명되었지만, 이와 반대로 불용어 사전과 반대되는 용어 사전을 이용하여 특정 키워드만을 선택할 수도 있을 것이다.

스팸 메일 인식부(540)는 스팸성 단어 저장부(550)에 수록된 스팸성 단어와 키워드 추출부(520)에서 선택된 키워드를 비교하고 비교 결과 사전 설정치 이상의 유사도, 예컨대, 90 % ~ 95 %의 유사도를 갖는 메일을 스팸 메일로서 인식한다. 스팸 메일 인식부(540)에서 유사도를 계산하는 방법은 다음과 같이 설명될 수 있다. 먼저, 메일의 문서를 구성하는 키워드와 그에 할당된 가중치를 메일 문서의 키워드 벡터로 하고, 스팸 용어 저장부(550)의 스팸성 단어와 그에 할당된 가중치를 스팸성 단어 벡터로 하여 두 벡터 사이의 유사도를 계산한다. 이때, 가중치는 어느 문서 내에서 특정 단어가 몇 번 출현하였는지를 나타내는 용어 빈도수(Term Frequency)와 특정 단어가 전체 문서에서 사용되고 있는 빈도수를 나타내는 역파일 빈도수(Inverse Frequency)를 이용하여 계산하며, 유사도 계산은 코사인 계수(Cosine Coefficient)를 사용할 수도 있다. 스팸 메일 인식부(540)는 스팸성 단어와 키워드 간의 유사도를 비교하여 해당 메일을 스팸 메일로 인식하며, 그 스팸 인식 결과를 메일 관리부(210)로 제공한다.

수신된 메일이 스팸 메일인 것으로 인식되는 경우에 사용자(140)(도 1 참조)는 도 6 및 도 7에서 설명하는 비와 같은 방법으로 스팸 메일을 처리할 수 있다.

도 6은 메일 서버(130)에서 수신된 각각의 메일에 대하여 스팸 메일 여부를 확인한 후에 사용자에게 일반 메일만을 전달해주는 경우를 예시한다.

도 6에서, 메일 서버(130)에서 수신된 메일에 대하여 스팸 메일의 여부를 판단한다(단계 S610). 판단 결과, 수신된 메일이 스팸 메일인 것으로 판단되면, 단계(S620)로 진행하고, 그렇지 않으면 단계(S630)로 진행한다.

단계(S620)에서, 메일 서버(130)는 해당 메일이 스팸 메일임을 나타내는 태크를 메일 저장부(220)에 저장된 메일에 기록하거나 그 스팸 메일을 폐기한다.

한편, 단계(S630)에서, 메일 서버(130)는 스팸 메일이라고 인식된 메일을 제외한 일반 메일을 사용자(140)에게 전송한다.

그 다음, 메일 서버(130)에서 사용자(140)가 메일 서버(130)에 로그인한 후 특별히 스팸 메일이라고 인식된 메일을 조회하고자 하는 요청이 있는지를 판단한다(단계 S640).

사용자(140)로부터 스팸 메일 조회 요청이 있으면, 메일 서버(130)는 단계(S650)로 진행하여 메일 저장부(220)로 이동하여 해당 사용자(140)가 조회를 요청한 메일이 스팸 메일로 표시되어 있는지를 조회하여 이를 사용자(140)에게 알려준다.

도 7은 메일 서버(130)에서 수신된 각각의 메일을 스팸 메일 여부를 확인하지 않은 채로 일단 사용자(140)에게 전달해준 다음 사후 처리하는 경우를 예시한다.

먼저, 단계(S710)에서 외부 메일 서버(110)로부터 수신된 메일은 메일 서버(130)에 의해 그대로 사용자(140)에게 전달된다.

그 다음, 메일 서버(130)는 사용자(140)에게 전달된 메일에 대하여 스팸 메일의 여부를 판단한다(단계 S720). 스팸 메일의 판단 결과, 사용자(140)에게 전달된 메일이 스팸 메일인 것으로 판단되면, 단계(S730)로 진행하여 사용자(140)에게 전달된 메일이 스팸 메일임을 알려준다. 사용자에게 스팸 메일이 전달되었음을 알리는 방법은 메일 저장부(220)에 저장되는 메일에 대하여 태크를 기록하여 해당 메일이 스팸 메일임을 알리는 것이다.

이후, 메일 서버(130)에 로그인한 사용자가 자신의 메일 목록에서 스팸 메일이 있음을 확인하고, 메일 서버(130)로 스팸 메일을 일괄 선택하여 줄 것을 요청하며, 메일 서버(130)는 사용자(140)로부터 스팸 메일 일괄 선택 요청이 있는지를 체크한다(단계 S740).

단계(S750)에서, 사용자(140)로부터 스팸 메일 일괄 선택 요청이 있음을 확인한 메일 서버(130)는 사용자(140)의 메일 목록에 포함된 스팸 메일을 자동 인식하고 스팸 메일 리스트를 작성하여 사용자(140)에게 제공한다.

이후, 메일 서버(130)는 스팸 메일 리스트를 확인한 사용자로부터의 스팸 메일의 삭제 요청에 따라 메일 저장부(220)에 저장되어 있는 스팸 메일을 일괄 삭제하거나 폐기한다.

본 발명은 상기한 실시예에 한정되지 않고, 본 발명의 기술적 요지를 벗어나지 않는 범위 내에서 다양하게 수정 및 변경 실시할 수 있음을 이 기술 분야에서 통상의 지식을 가진 자라면 누구나 이해할 수 있을 것이다.

발명의 효과

본 발명에 따르면, 스팸 메일 필터링을 통하여, 상업성 및 광고성 내용을 포함하는 메일이 사용자에게 무분별하게 전달되는 것을 사전에 차단함으로써, 사용자들의 메일 시스템을 이용하는 관리적인 수고를 줄여 줄 수 있다.

본 발명의 스팸 필터링 기법을 이용하여, 사용자 입장에서 정보성 및 중요도가 낮은 메일들을 대상으로 메일 인식, 차단 및 관리 등의 편의성을 제공함으로써, 사용자의 메일 관리 작업의 효율성을 높여줄 수

있다.

본 발명의 대량 메일에 대한 정확한 파악을 통하여, 대량 메일을 발송하는 이메일 발송자들에 대한 정확한 정보 및 발송 현황 등을 파악하고, 이러한 명확한 기준을 바탕으로 해당 발송자에 대하여 필요한 조치를 취할 수 있다.

또한, 대량의 스팸 메일을 발송하는 발송자를 파악하고 이에 대한 조치를 행함으로써 불필요한 인터넷 트래픽을 감소시키고, 이로 인하여 인터넷에서 데이터 전송 속도를 증대하는데 기여할 수 있다.

(57) 청구의 범위

청구항 1

메일 서버에서 스팸 메일을 필터링하는 방법에 있어서,

외부로부터 수신된 각각의 메일에 대하여 유사 메일의 발송 건수를 기반으로 상기 스팸 메일을 판단하는 것을 특징으로 하는 스팸 메일 필터링 방법.

청구항 2

제 1 항에 있어서,

상기 스팸 메일 판단 단계는,

- (a) 외부로부터 수신된 각각의 메일에 포함된 문자 중에서 사용 빈도수가 높은 N(N은 자연수) 개의 문자를 상기 메일을 대표하는 대표 문자열로 추출하는 대표 문자 추출 단계;
- (b) 상기 대표 문자열이 동일한 메일들의 개수를 상기 대표 문자열별로 누적하는 메일 누적 단계; 및
- (c) 상기 대표 문자열별로 누적된 메일들의 누적 개수가 기 설정 개수 이상인 메일들을 상기 유사 메일로 분석하는 유사 메일 분석 단계; 및
- (d) 상기 유사 메일로 분석된 메일들을 상기 스팸 메일로서 판단하는 스팸 메일 판단 단계

를 포함하는 것을 특징으로 하는 스팸 메일 필터링 방법.

청구항 3

제 1 항에 있어서,

상기 유사 메일 분석 단계 (c)는,

- (c-1) 상기 각각의 유사 메일에 포함된 전체 문자의 사용 빈도수를 계산하는 단계;
- (c-2) 상기 각각의 유사 메일에서 각 문자의 사용 빈도수의 평균값을 계산하는 단계;
- (c-3) 상기 각 문자의 사용 빈도수와 상기 평균값을 비교하는 단계;
- (c-4) 상기 비교 결과, 상기 각 문자의 사용 빈도수와 상기 평균값과의 유사도가 사전 설정치를 초과하는 메일을 상기 유사 메일로 인식하는 단계

를 구비함으로써, 상기 유사 메일로 분류된 메일들 중에서 상기 대표 문자열이 동일하지만 메일의 내용이 일부 바뀐 메일을 상기 유사 메일로 인식하는 것을 특징으로 하는 스팸 메일 필터링 방법.

청구항 4

제 1 항에 있어서,

상기 유사 메일 분석 단계 (c)는,

- (c-1) 상기 각각의 유사 메일에 포함된 전체 문자의 사용 빈도수를 계산하는 단계;
- (c-2) 상기 각각의 유사 메일에서 각 문자의 사용 빈도수의 평균값을 계산하는 단계;
- (c-3) 상기 각 문자의 사용 빈도수와 상기 평균값을 비교하는 단계; 및
- (c-4) 상기 비교 결과, 상기 각 문자의 사용 빈도수와 상기 평균값과의 유사도가 사전 설정치를 초과하지 않는 메일을 상기 유사 메일에서 제외시키는 단계

를 구비함으로써, 상기 유사 메일로 분류된 메일들 중에서 상기 대표 문자열이 동일하지만 메일의 내용이 유사하지 않은 메일을 제외한 메일을 상기 유사 메일로 인식하는 것을 특징으로 하는 스팸 메일 필터링 방법.

청구항 5

메일 서버에서 스팸 메일을 필터링하는 방법에 있어서,

- (a) 외부로부터 수신된 각각의 메일에 포함된 문자 중에서 사용 빈도수가 높은 N(N은 자연수) 개의 문자를 상기 메일을 대표하는 대표 문자열로 추출하는 대표 문자 추출 단계;
- (b) 상기 대표 문자열이 동일한 메일들의 개수를 상기 대표 문자열별로 누적하는 메일 누적 단계;
- (c) 상기 대표 문자열별로 누적된 메일들의 누적 개수가 기 설정 개수 이상인 메일들을 대량 메일로 분석하는 대량 메일 분석 단계; 및

(d) 상기 대량 메일로 분류된 각각의 메일에 포함된 복수개의 키워드와 기설정 키워드를 비교하여 기설정 사전 설정치 이상의 유사도를 갖는 메일을 상기 스팸 메일로 인식하는 스팸 메일 분석 단계를 포함하는 것을 특징으로 하는 메일 서버의 스팸 메일 필터링 방법.

청구항 6

제 5 항에 있어서.

상기 스팸 메일 필터링 방법은,

(f) 상기 대량 메일로 분류된 상기 메일 중에서 상기 대표 문자열이 동일하지만 메일의 내용이 유사하지 않은 메일을 제외시키는 비유사 메일 제외 단계

를 더 포함하는 것을 특징으로 하는 스팸 메일 필터링 방법.

청구항 7

제 6 항에 있어서.

상기 비유사 메일 제외 단계 (f)는

(f-1) 상기 각각의 메일에 포함된 전체 문자의 사용 빈도수를 계산하는 단계;

(f-2) 상기 각각의 메일에서 각 문자의 사용 빈도수의 평균값을 계산하는 단계;

(f-3) 상기 각 문자의 사용 빈도수와 상기 평균값을 비교하는 단계;

(f-4) 상기 비교 결과, 상기 각 문자의 사용 빈도수와 상기 평균값과의 유사도가 사전 설정치를 초과하지 않는 메일을 상기 대량 메일에서 제외시키는 단계

를 구비하는 것을 특징으로 하는 스팸 메일 필터링 방법.

청구항 8

제 7 항에 있어서.

상기 단계 (f-4)는 상기 대량 메일로 분류된 메일들 중에서 상기 대표 문자열이 동일하지만 메일의 내용이 일부 바뀐 메일을 상기 대량 메일로 인식하는 것을 특징으로 하는 스팸 메일 필터링 방법.

청구항 9

제 5 항에 있어서.

상기 스팸 메일 분석 단계 (d)는,

(d-1) 상기 메일의 내용에 포함된 문자열에 대하여 형태소 단위로 분석하여 추출된 복수개의 단어를 키워드 가능 후보로 선택하는 단계;

(d-2) 상기 선택된 키워드 가능 후보 중에서 불용어를 제외한 키워드 가능 후보를 상기 메일을 대표하는 키워드로서 선택하는 단계;

(d-3) 상기 선택된 키워드를 기설정 스팸성 정보와 비교하는 단계;

(d-4) 상기 비교 결과, 상기 선택된 키워드와 상기 기설정 스팸성 정보와의 유사도가 사전 설정치 이상일 때, 상기 메일을 스팸 메일로서 인식하는 단계

를 구비하는 것을 특징으로 하는 스팸 메일 필터링 방법.

청구항 10

상업성 또는 광고성 스팸 메일을 필터링하는 장치에 있어서,

외부로부터 수신된 메일들 중에서 스팸 메일 인식 결과에 따라 상기 수신된 각각의 메일에 대한 스팸 메일의 여부를 사용자에게 알려주는 메일 관리부; 및

상기 메일 관리부에서 수신된 각각의 메일에 포함된 내용을 분석하여 내용이 유사한 메일들을 분류하고, 분류된 메일들이 기설정 개수 이상으로 대량인 유사 메일들을 상기 스팸 메일로 분석하는 대량 메일 분석부

를 포함하는 것을 특징으로 하는 스팸 메일 필터링 장치.

청구항 11

제 10 항에 있어서.

상기 대량 메일 분석부는

상기 각 문자의 사용 빈도수를 계산하고, 상기 사용 빈도수가 가장 많은 N 개의 문자를 상기 메일을 대표하는 대표 문자열로 추출하는 대표 문자 추출부;

상기 대표 문자열이 동일한 메일들을 분류하는 동일 메일 분류부;

상기 동일 메일 분류부에 의해 분류된 메일을 상기 대표 문자열별로 누적하는 메일 관리 테이블; 및

상기 메일 관리 테이블을 참조하여 상기 대표 문자열별로 분류된 메일들의 누적 개수가 상기 기설정 개수

이상인 메일들을 상기 유사 메일로 판별하는 대량 메일 판별부
를 구비하는 것을 특징으로 하는 스팸 메일 필터링 장치.

청구항 12

제 10 항에 있어서,

상기 스팸 메일 필터링 장치는,

상기 대량 메일 분석부에 의해 유사 메일로 분류된 상기 메일들 중에서 상기 대표 문자가 동일하되 메일의 내용이 유사하지 않은 메일을 제외시키는 비유사 메일 분석부를 더 포함하며.

상기 비유사 메일 분석부는,

상기 각각의 유사 메일에 포함된 전체 문자의 사용 빈도수를 계산하는 문자 빈도수 계산부;

상기 문자 빈도수 계산부에 의해 계산된 상기 유사 메일에서 각 문자의 사용 빈도수의 평균값을 계산하는 빈도수 평균 계산부; 및

상기 각 문자의 사용 빈도수와 상기 평균값을 비교한 유사도가 사전 설정치를 초과하지 않는 메일을 상기 동일 메일에서 제외시키는 빈도수 평균 비교부

를 구비하는 것을 특징으로 하는 스팸 메일 필터링 장치.

청구항 13

상업성 또는 광고성 스팸 메일을 필터링하는 장치에 있어서,

외부로부터 수신되는 각각의 메일을 스팸 메일 인식 결과에 따라 스팸 메일로서 관리하는 메일 관리부;

상기 메일 관리부에서 수신된 각각의 메일에 포함된 문자 중에서 사용 빈도수가 높은 N(N은 자연수) 개의 문자를 상기 메일을 대표하는 대표 문자열로 추출하고, 상기 추출된 대표 문자열이 동일한 메일들이 사전 설정치 이상인 메일들을 대량 메일로 분류하는 대량 메일 분석부; 및

상기 대량 메일로 분류된 메일에 포함된 단어들 중에서 그 메일을 대표하는 키워드와 기설정 스팸성 키워드를 비교하여 사전 설정치 이상의 유사도를 갖는 메일을 스팸 메일로 인식하고 상기 메일 관리부로 상기 스팸 메일 인식 결과를 제공하는 스팸 메일 분석부

를 포함하는 것을 특징으로 하는 스팸 메일 필터링 장치.

청구항 14

제 13 항에 있어서,

상기 대량 메일 분석부는

상기 각 문자의 사용 빈도수를 계산하고, 상기 사용 빈도수가 가장 많은 N 개의 문자를 상기 대표 문자열로 추출하는 대표 문자 추출부;

상기 대표 문자열이 동일한 메일들을 분류하는 동일 메일 분류부;

상기 동일 메일 분류부에 의해 분류된 메일을 상기 대표 문자열별로 누적하는 메일 관리 테이블; 및

상기 메일 관리 테이블을 참조하여 상기 대표 문자열에 속하는 메일의 누적 개수가 기설정 개수 이상인 메일들을 상기 대량 메일로 판별하는 대량 메일 판별부

를 구비하는 것을 특징으로 하는 스팸 메일 필터링 장치.

청구항 15

제 13 항에 있어서,

상기 스팸 메일 분석부는

상기 대량 메일로 분류된 각각의 메일의 내용에 포함된 문자열에 대하여 형태소 단위의 분석을 수행하여 분석된 적어도 하나의 단어를 추출하고, 상기 추출된 단어를 키워드 가능 후보로 제공하는 형태소 분석부;

상기 형태소 분석부에 의해 제공된 상기 키워드 가능 후보 중에서 상기 메일의 문서를 대표하는 키워드를 선택하는 키워드 추출부; 및

상기 키워드 추출부에서 추출된 키워드를 기설정 스팸성 정보와 비교하여 사전 설정치 이상의 유사도를 갖는 메일을 스팸 메일로서 인식하고 상기 스팸 메일 인식 결과를 상기 메일 관리부로 제공하는 스팸 메일 인식부

를 구비하는 것을 특징으로 하는 스팸 메일 필터링 장치.

청구항 16

제 15 항에 있어서,

상기 스팸 메일 분석부는 사용되지 않는 불용어를 수록한 불용어 키워드 저장부를 더 구비하며,

상기 키워드 추출부는 상기 키워드 가능 후보들 중에서 상기 불용어 키워드 저장부에 수록되지 않은 모든 키워드 가능 후보를 상기 키워드로서 추출하는 것을 특징으로 하는 스팸 메일 필터링 장치.

청구항 17

제 15 항에 있어서,

상기 스팸 메일 분석부는 상기 기설정 스팸성 정보를 수록하고 있는 스팸 용어 저장부를 더 구비하며.

상기 스팸 메일 인식부는 상기 키워드 및 상기 기설정 스팸성 정보에 가중치를 할당하여 상기 키워드와 상기 기설정 스팸성 정보와의 유사도를 계산하는 것을 특징으로 하는 스팸 메일 필터링 장치.

청구항 18

제 17 항에 있어서,

상기 가중치는 용어 빈도수(Term Frequency) 및 역파일 빈도수(Inverse Frequency)의 값이며. 상기 유사도는 코사인 계수(Cosine Coefficient)로 계산되는 것을 특징으로 하는 메일 서버의 스팸 메일 필터링 장치.

청구항 19

제 13 항에 있어서,

상기 스팸 메일 필터링 장치는,

상기 대량 메일 분석부에 의해 분류된 상기 유사 메일들 중에서 상기 대표 문자가 동일하되 메일의 내용이 유사하지 않은 메일을 제외시키는 비유사 메일 분석부를 더 포함하며.

상기 비유사 메일 분석부는,

상기 각각의 유사 메일에 포함된 전체 문자의 사용 빈도수를 계산하는 문자 빈도수 계산부;

상기 문자 빈도수 계산부에 의해 계산된 상기 유사 메일에서 각 문자의 사용 빈도수의 평균값을 계산하는 빈도수 평균 계산부; 및

상기 각 문자의 사용 빈도수와 상기 평균값을 비교한 유사도가 사전 설정치를 초과하지 않는 메일을 상기 유사 메일에서 제외시키는 빈도수 평균 비교부

를 구비하는 것을 특징으로 하는 스팸 메일 필터링 장치.

청구항 20

제 13 항에 있어서,

상기 메일 관리부는 상기 스팸 메일 분석부에서 제공된 상기 스팸 메일 분석 결과에 따라 상기 스팸 메일로 판별된 메일을 제외한 메일들만을 상기 사용자에게 전송하는 것을 특징으로 하는 스팸 메일 필터링 장치.

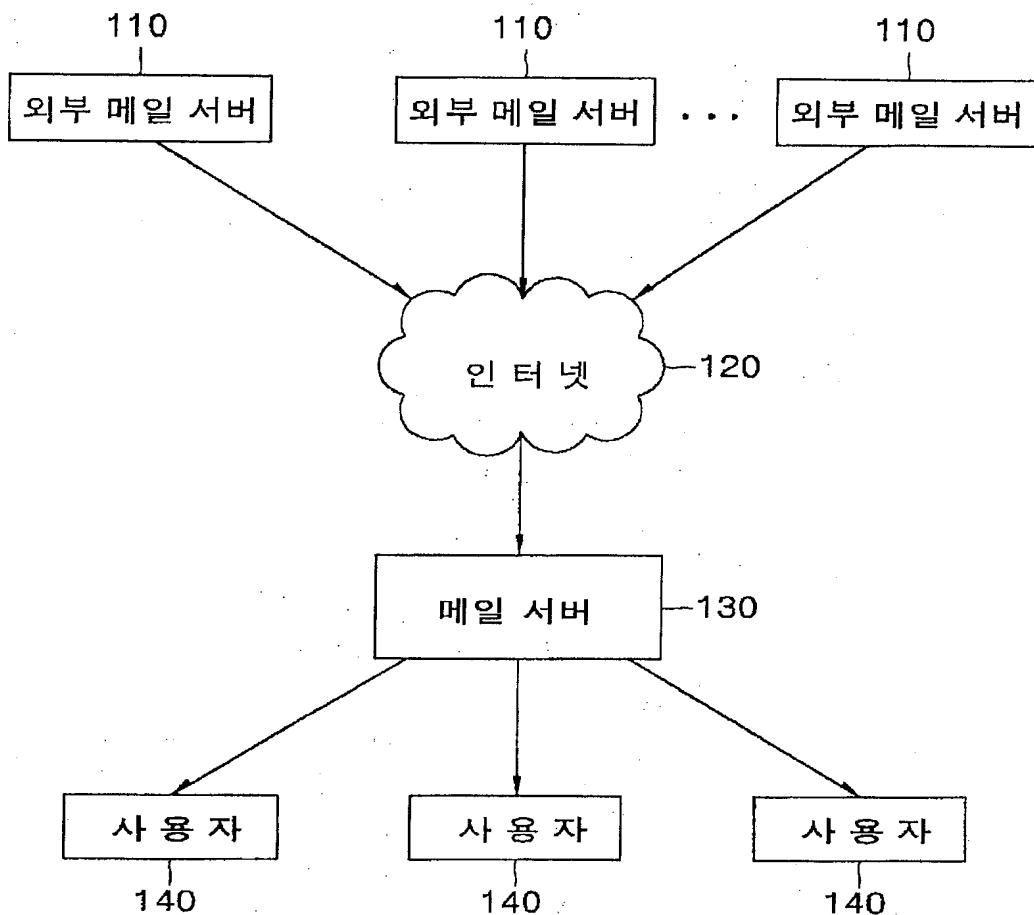
청구항 21

제 13 항에 있어서,

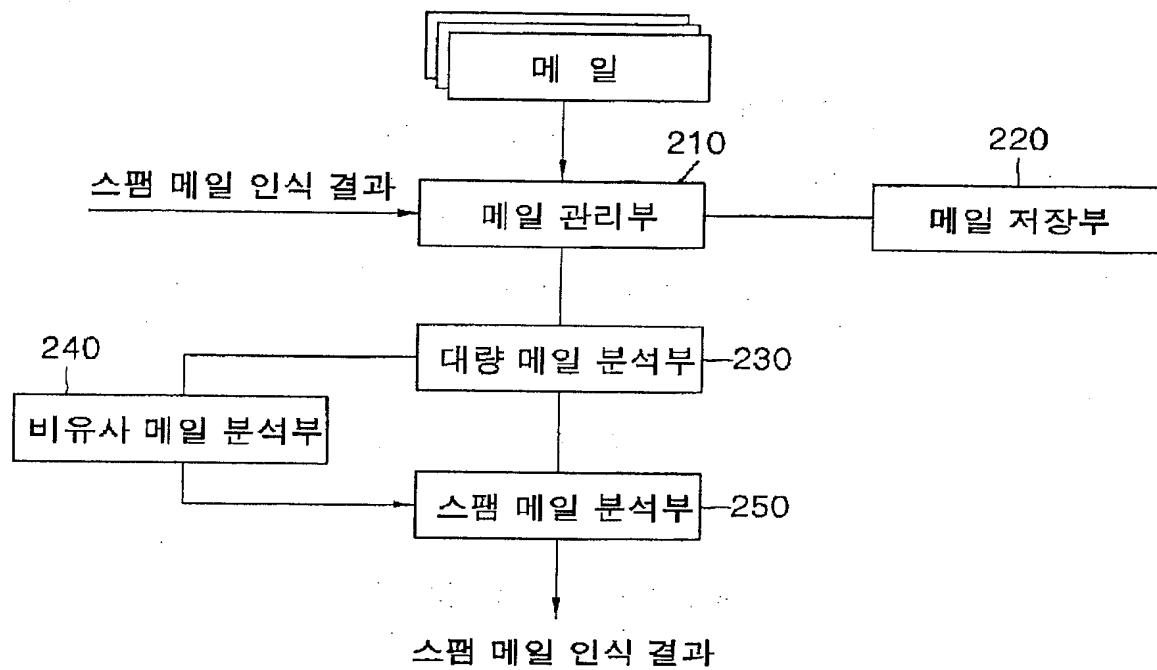
상기 메일 관리부는 상기 수신된 메일을 그대로 상기 사용자에게 제공하고, 상기 스팸 메일 분석부에서 제공된 상기 스팸 메일 분석 결과를 상기 사용자에게 통보하는 것을 특징으로 하는 스팸 메일 필터링 장치.

도면

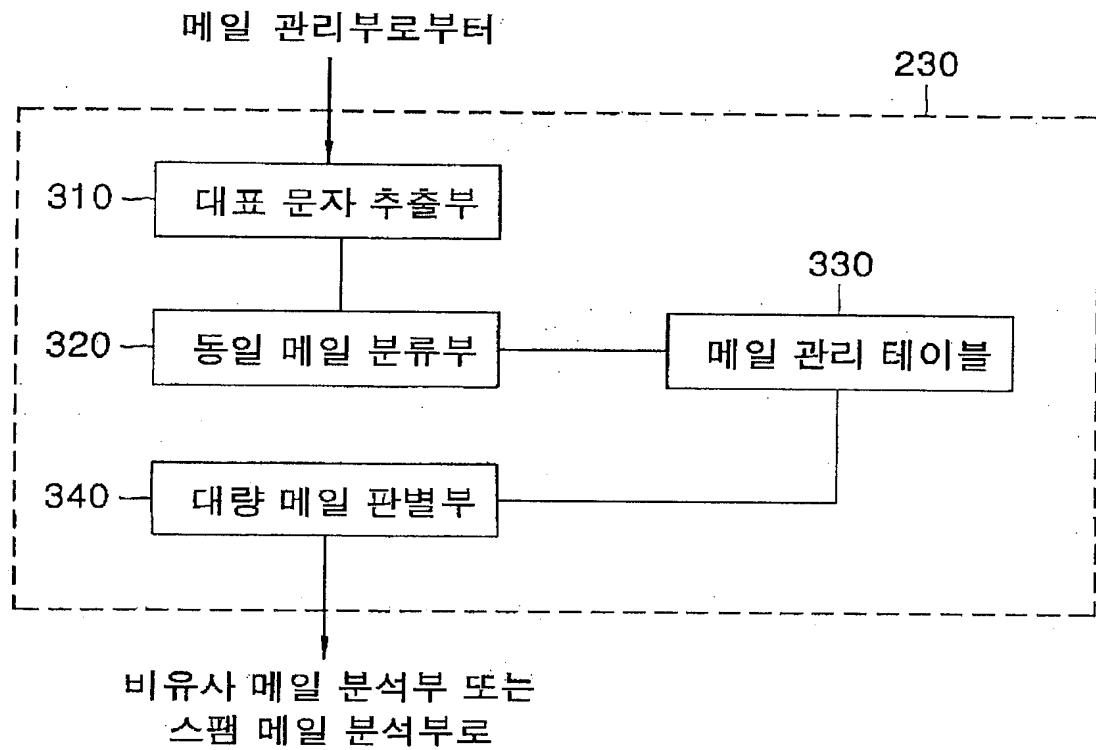
도면1



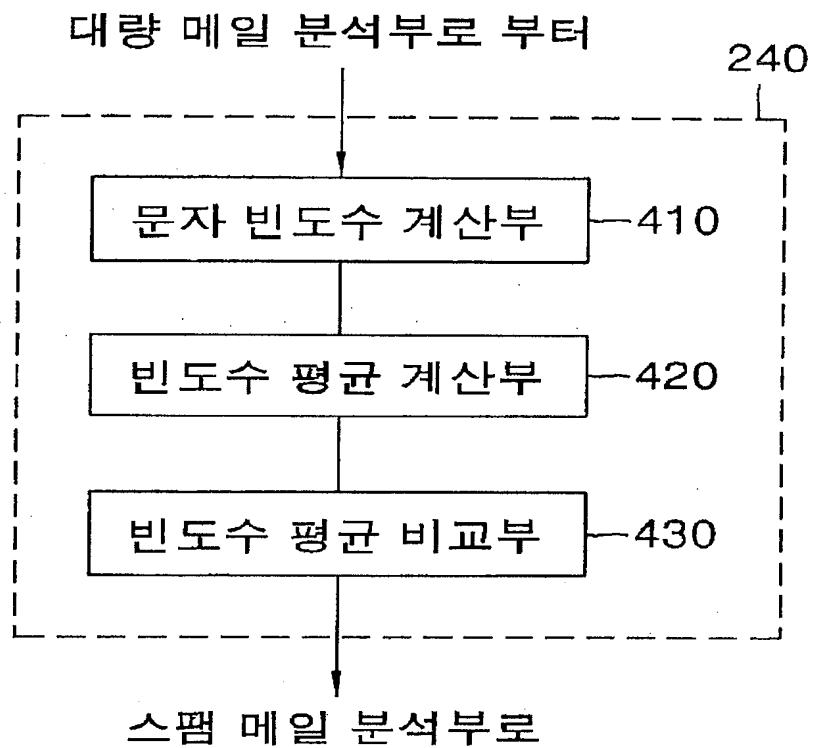
도면2

130

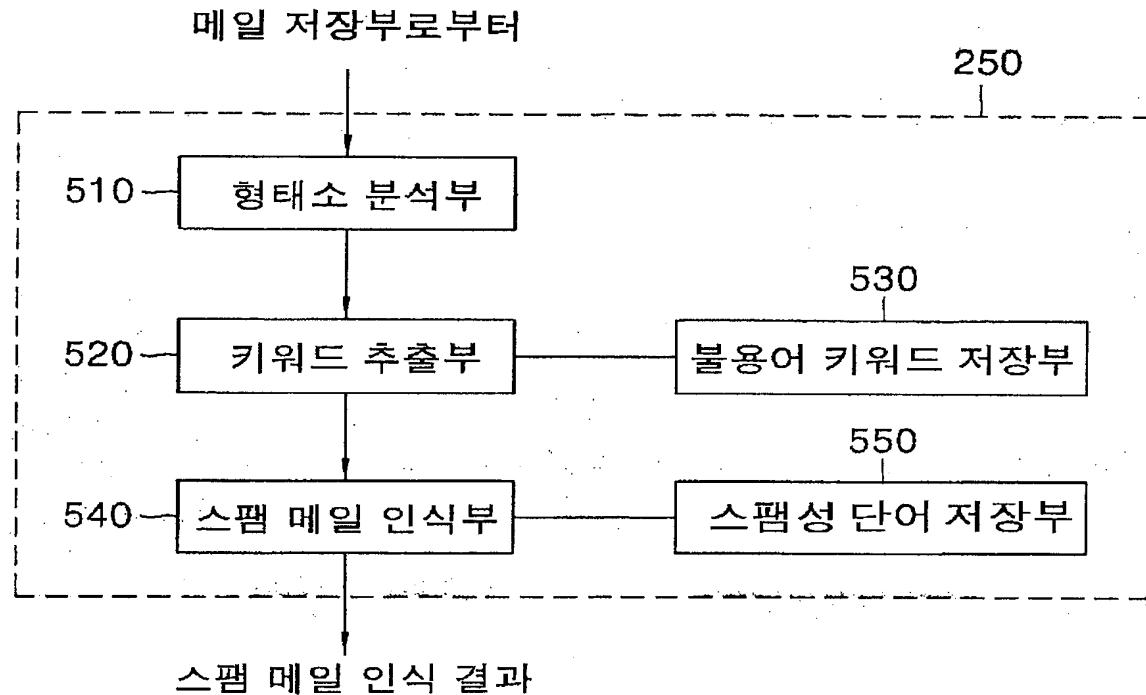
도면3



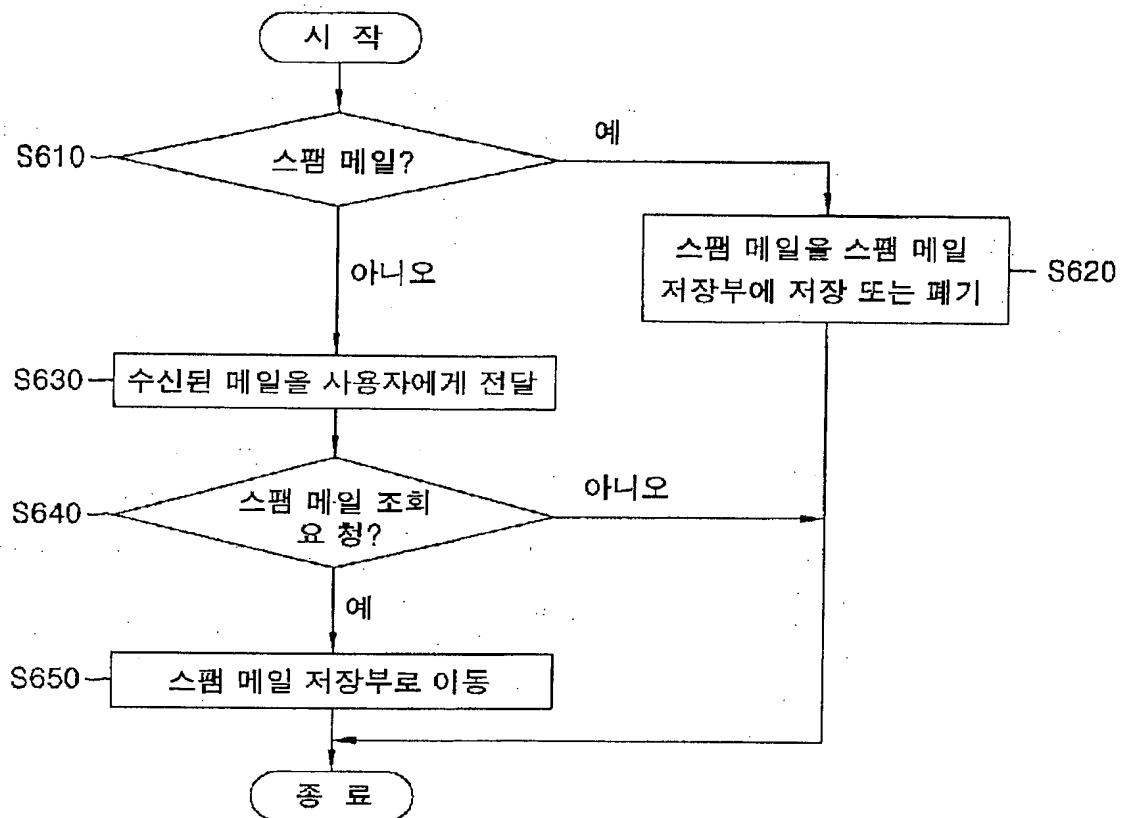
도면4



도면5



도면6



도면7

